Chapitre 3 : compléments sur le modèle linéaire

Voici les hypothèses du modèle linéaire. On veut prédire une variable Y à l'aide de plusieurs covariables numériques $X_1, \ldots, X_p$. Un **individu** de la population est donc représenté par le vecteur $(X_1, \ldots, X_p, Y)$. On suppose que

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

où $\varepsilon \perp\!\!\!\perp (X_1, \ldots, X_p)$, $\mathbb{E}(\varepsilon) = 0$ et $\mathrm{Var}(\varepsilon) = \sigma^2$. Les paramètres de ce modèle sont $\beta_0, \ldots, \beta_p, \sigma^2$.

Sous l'hypothèse gaussienne, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

1 Autres méthodes d'inférence

1.1 Problèmes avec l'estimateur des moindres carrés

Dans toute cette partie, on supposera σ^2 connu. On rappelle que **l'estimateur des moindres carrés** est donné par

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Lorsque le nombre de covariables est important, et ou que celles-ci sont corrélées, l'estimateur devient instable. En effet, si les covariables sont centrées, toutes les colonnes de $\mathbf{X}$ sont centrées, sauf celle constante égale à 1 si on veut inclure β_0 dans le modèle. Le coefficient j,k de la matrice $\mathbf{X}'\mathbf{X}$ est alors

$$\sum_{i=1}^{n}(X_{i,j} - \bar{X}_j)(X_{i,k} - \bar{X}_k) = (n-1)\widehat{\sigma}(X_j, X_k).$$

Si la corrélation est importante, l'inversion de $\mathbf{X}'\mathbf{X}$ est instable. Et, par

exemple

$$\begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}^{-1} \approx \begin{pmatrix} 2.78 & -2.22 \\ -2..2 & 2.78 \end{pmatrix}, \qquad \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}^{-1} \approx \begin{pmatrix} 5.26 & -4.74 \\ -4.74 & 5.26 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}^{-1} \approx \begin{pmatrix} 10.26 & -9.74 \\ -9.74 & 10.26 \end{pmatrix}$$

Toute erreur d'estimation dans la matrice de covariance due à l'échantillon des $X_{i,j}$ peut faire varier $(\mathbf{X}'\mathbf{X})^{-1}$ fortement.

L'estimateur des moindres carrées est donc non biaisé, mais de variance importante. Pour obtenir des estimateurs plus efficaces, il faut utiliser des estimateurs biaisés.

1.2 Estimation bayésienne

Le principe de l'estimation bayésienne est de coder l'information dont on dispose sur β **avant d'observer les données** sous forme d'une **loi a priori**. Puis de

calculer une loi a posteriori, qui représente l'information après avoir observé les données.

Pour éviter que les valeurs de β n'explosent, on peut essayer de choisir une loi a priori qui favorise des valeurs de β autour de 0. Evidemment, l'unité dans laquelle la variable X_j influe sur la valeur de β_j :

$$\beta_j X_j = (\lambda \beta_j)\left(\frac{X_j}{\lambda}\right).$$

Autrement dit, quand on divise X_j par λ, la valeur de β_j est multipliée par λ.

Dans toute la suite de cette partie 1, on suppose que les variables X_j sont centrées et réduites.

L'a priori

$$\pi(\beta_0) \propto 1, \quad \beta_{1:p} \sim \mathcal{N}_p(0, \tau^2 I_d)$$

favorise les $\beta_{1:p}$ proches de 0. Comme toutes les covariables sont réduites, le fait que la variance a priori de β_i, $i = 1, \ldots, p$ soit identique, égale à τ^2 quelque soit i permet de tirer toutes les estimations de β_i vers 0 avec la même force.

Le modèle linéaire gaussien suppose que

$$\left[\mathbf{Y}\middle|\mathbf{X},\boldsymbol{\beta}\right] \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2 I_n)$$

Dans ce cas, la loi a posteriori est donnée par

$$\pi\left(\boldsymbol{\beta}\middle|\mathbf{Y},\mathbf{X}\right) \propto \exp\left(-\frac{1}{2\sigma^2}\left\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right\|^2 - \frac{1}{2\tau^2}\left\|\beta_{1:p}\right\|^2.\right)$$

La log-posterior est donc

$$\log\pi\left(\boldsymbol{\beta}\middle|\mathbf{Y},\mathbf{X}\right) = -\frac{1}{2\sigma^2}\left\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\right\|^2 - \frac{1}{2\tau^2}\left\|\beta_{1:p}\right\|^2 + \text{Cte.}$$

L'**estimateur du maximum a posteriori** (ou MAP) est celui qui maximie la log-posterior, donc qui minimise la fonction

$$\mathbf{b} \mapsto \left\|\mathbf{Y}-\mathbf{X}\mathbf{b}\right\|^2 + \frac{\sigma^2}{\tau^2}\left\|\mathbf{b}_{1:p}\right\|^2.$$

On voit ici que l'on se retrouve à minimiser le critère des moindres carrés, mais **pénalisé par la norme** $\left\|\mathbf{b}_{1:p}\right\|^2$. Cela revient à chercher un **compromis** entre une valeur de $\mathbf{b}$ qui minimise les moindres carrés, et une valeur qui minimise la norme. La valeur de τ règle se compromis :

— quand τ est grand, la loi a priori favorise peu les valeurs de $\beta_{1:p}$ autour de 0, et, de fait, la pénalisation du critère des moindres carrés est faible,
— quand τ est faible, l'a priori est fort, et la pénalisation est forte.

Il existe évidemment **d'autres lois a priori classiques** pour traiter le cas du modèle linéaire, comme les lois a priori de dites g-**prior de Zellner**, qui proposent

$$\left[\boldsymbol{\beta}\middle|\mathbf{X}\right] \sim \mathcal{N}\left(0, g\sigma^2\mathbf{X}'\mathbf{X}\right).$$

Cette loi a priori ne favorise pas les valeurs de $\boldsymbol{\beta}$ vers 0 en cas de forte corrélations entre celles-ci. Dans ce cas, la loi a posteriori est aussi explicite et est donnée par

$$\left[\boldsymbol{\beta}\middle|\mathbf{Y},\mathbf{X}\right] \sim \mathcal{N}\left(\frac{g}{g+1}\widehat{\boldsymbol{\beta}}_{\mathsf{ML}}, \frac{\sigma^2 g}{g+1}(\mathbf{X}'\mathbf{X})^{-1}\right)\ldots$$

Quand $g = n$, cette loi a posteriori permet de retrouver des **résultats similaires à la statistique fréquentielle avec le maximum de vraisemblance**. Elle est plutôt utilisée pour faire du choix de modèle bayésien, voir plus tard dans le cours. . .

1.3 Estimateur ridge

Les **estimateurs ridges** sont exactement les estimateurs qui minimisent l'un des critères des moindres carrés pénalisés par la norme $\|\beta_{1:p}\|^2$. Ils permettent de **régler les problèmes de corrélations trop fortes dans les covariables.**

Remarquons que, comme les variables X_j sont centrées (et réduites), on a

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \mathbf{X}'_{-0}\mathbf{X}_{-0} & \\ 0 & & & \end{pmatrix}$$

où $\mathbf{X}'_{-0}\mathbf{X}_{-0}$ est la matrice de corrélation des covariables.

Par définition, les estimateurs ridge s'écrivent sous la forme

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} = \left(\mathbf{X}'\mathbf{X} + \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \lambda^2 I_p & \\ 0 & & & \end{pmatrix} \right)^{-1} \mathbf{X}'\mathbf{Y}, \quad (\lambda \geq 0).$$

Lorsque $\lambda = 0$, on retrouve l'estimateur des moindres carrés.

Ce sont les **solutions du problème d'optimisation du critère des moindres carrés pénalisés**

$$\mathbf{b} \mapsto \left\|\mathbf{Y} - \mathbf{X}\mathbf{b}\right\|^2 + \lambda^2 \left\|\mathbf{b}_{1:p}\right\|^2.$$

Notons qu'il s'agit exactement du critère qui définissait le MAP lorsque la loi a priori sur $\beta_{1:p}$ est gaussienne centrée en 0, en posant $\lambda^2 = \sigma^2/\tau^2$.

La valeur de λ fixe la force de la correction. Elle doit être choisi en fonction du jeu de données, par validation croisée (voir plus bas). Notons, par exemple, que

$$\begin{pmatrix} 1.2 & 0.8 \\ 0.8 & 1.2 \end{pmatrix}^{-1} \approx \begin{pmatrix} 1.5 & -1.0 \\ -1.0 & 1.5 \end{pmatrix}, \qquad \begin{pmatrix} 1.2 & 0.9 \\ 0.9 & 1.2 \end{pmatrix}^{-1} \approx \begin{pmatrix} 1.90 & -1.43 \\ -1.43 & 1.90 \end{pmatrix},$$

$$\begin{pmatrix} 1.2 & 0.95 \\ 0.95 & 1.2 \end{pmatrix}^{-1} \approx \begin{pmatrix} 2.23 & -1.77 \\ -1.77 & 2.23 \end{pmatrix}.$$

On voit ici, avec $\lambda^2 = 0.2$ que l'on se tromper dans l'estimation de la covariance (qui vaut 0.8 dans cet exemple), et utiliser des estimations de l'ordre de 0.9 ou 0.95 sans que l'inverse n'explose (à comparer avec les calculs numériques de 1.1).

Une autre façon de comprendre ces estimateurs est de regarder **les valeurs propres de $\mathbf{X}'_{-0}\mathbf{X}_{-0}$, la matrice de corrélation des covariables**. Rappelons que la matrice de corrélation $\mathbf{X}'_{-0}\mathbf{X}_{-0}$ est symétrique, semi-définie positive. Elle est donc diagonalisable, et toutes ses valeurs propres sont ≥ 0. Dire que les covariables sont fortement corrélées revient à dire que certaines de ces valeurs propres sont proches de 0. En ajoutant λ^2 à toutes les valeurs propres, on les éloigne de 0.

Une dernière façon de voir ces estimateurs est de comprendre le critère des moindres carrés pénalisés

$$\mathbf{b} \mapsto \left\|\mathbf{Y} - \mathbf{X}\mathbf{b}\right\|^2 + \lambda^2 \left\|\mathbf{b}_{1:p}\right\|^2.$$

comme le lagrangien du problème d'optimisation du critère des moindres carrés sous la contrainte

$$\left\|\beta_{1:p}\right\|^2 \leq r^2.$$

La valeur de r^2 dépend bien sûr de la valeur de λ^2 : r^2 est grand quand λ^2 est petit, et réciproquement. Cette contrainte empêche l'estimateur des moindres carrés d'être trop grand. . . À nouveau, on voit qu'il faut rendre toutes les valeurs de $\boldsymbol{\beta}_i$ comparable donc centrer **réduire** les covariables avant d'appliquer ridge.

1.4 Pénalisation Lasso

La pénalisation Lasso est une autre pénalisation du critère des moindres carrés qui permet de forcer l'annulation de certains effets β_i. Initialement proposé par Tibshirani en 1996, Lasso veut dire « *least absolute shrinkage and selection operator* ». Les estimateurs Lasso sont définis par

$$\widehat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} = \operatorname{argmin}_{\mathbf{b}} \left(\left\| \mathbf{Y} - \mathbf{X}\mathbf{b} \right\|^2 + \lambda \left\| \mathbf{b}_{1:p} \right\|_1 \right)$$

où

$$\left\| \mathbf{b}_{1:p} \right\|_1 = \sum_{i=1}^{p} |b_i|$$

est la norme L^1 ou de Manhattan de $\mathbf{b}_{1:p}$. Comme pour la régression ridge, on peut voir ce critère pénalisé comme un lagrangien du problème d'optimisation du critère des moindres carrés sous la contrainte

$$\left\| \mathbf{b}_{1:p} \right\|_1 \leq r.$$

Comme la norme de Manhattan a des « boules » de rayon r qui sont des « losanges » avec des angles au niveau des axes, ce problème d'optimisation sous contraintes rend certaines coordonnées b_i nulles, voir Figure 1. Contrairement à

l'estimateur ridge, il n'y a pas de formule explicite pour la solution de ce problème d'optimisation.

À nouveau, il existe autant d'estimateur Lasso que de valeurs de λ possibles. Il faudra trouver une façon de choisir cette valeur, par validation croisée par exemple.

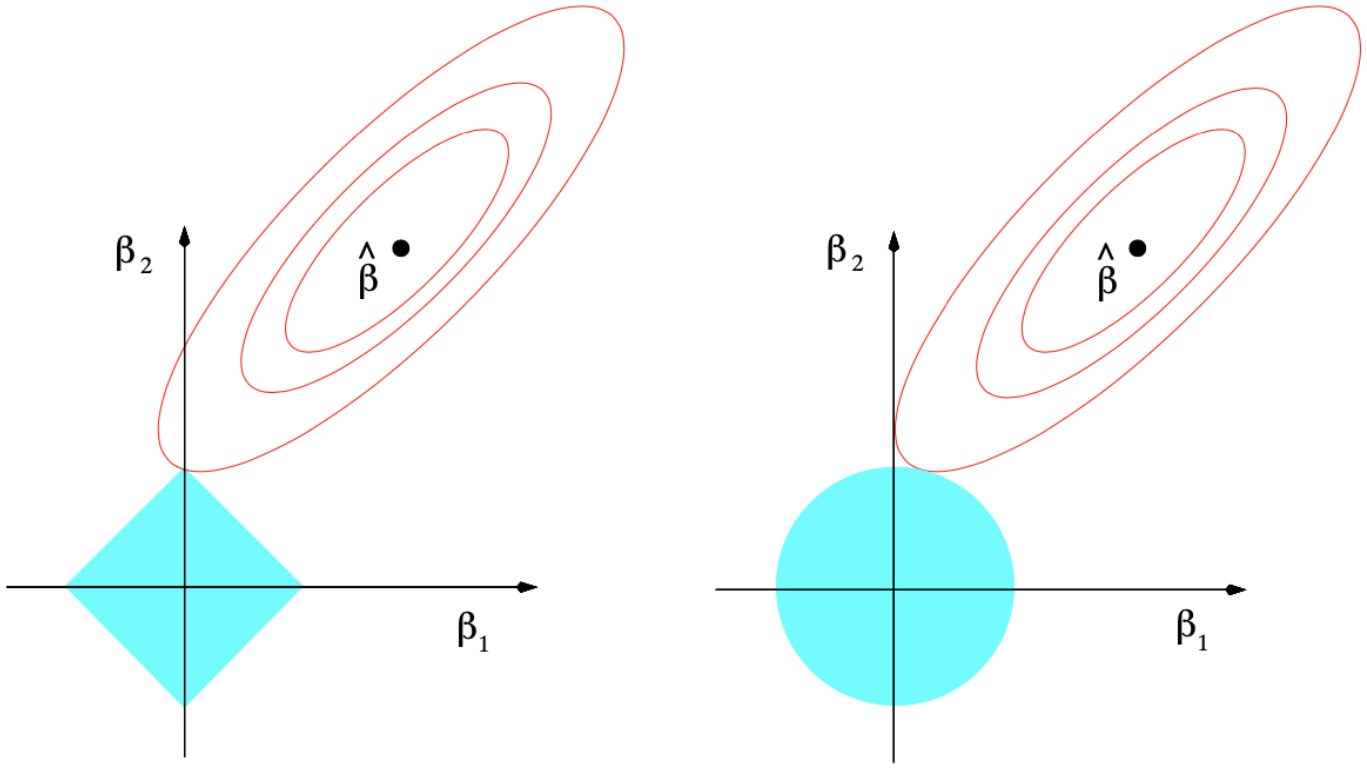

Figure 1 – **Estimation Lasso (à gauche) et ridge (à droite).** On s'intéresse à un problème où $\beta_0 = 0$. Les ellipses rouges sont les lignes de niveau du critère des moindres carrés, minimal en $\widehat{\beta}$, le maximum de vraisemblance.
On voit ici que l'estimateur Lasso va annuler $\widehat{\beta}_1^{\text{Lasso}}$, alors que l'estimateur ridge conserve une coordonnée $\widehat{\beta}_1^{\text{Ridge}}$ non nulle.

2 Validation croisée

L'idée de la validation croisée est d'étudier l'erreur de généralisation d'une estimation sur de nouvelles données, en utilisant les données comme de nouvelles données. Autrement dit, on va

- enlever des observations,
- ajuster le modèle sur les observations conservées, puis finalement
- faire des prédications sur les observations enlevées au début.

On suppose ici que l'on veut comparer des estimateurs $\widehat{\boldsymbol{\beta}}_\lambda$ qui dépendent d'un paramètres λ à estimer. Comme dans les estimateurs ridge et lasso par exemple.

Fixons un jeu de données $\mathbf{X}, \mathbf{Y}$ sur lequel les $\widehat{\boldsymbol{\beta}}_\lambda$ sont calculés. Et introduisons un nouvel individu $\widetilde{\mathbf{x}}$, $\widetilde{Y}$, indépendant du jeu de données. Si on veut prédire $\widetilde{Y}$ avec les covariables $\widetilde{\mathbf{x}}$ de l'individu, on utilise

$$\widehat{Y}_\lambda = \widetilde{\mathbf{x}}\widehat{\boldsymbol{\beta}}_\lambda.$$

L'erreur quadratique de prédiction est donnée par

$$SSE_{\text{gen}}(\lambda) = \mathbb{E}\left[(\widetilde{Y} - \widehat{Y}_\lambda)^2 \middle| \mathbf{X}, \mathbf{Y}\right] = \mathbb{E}\left[(\widetilde{Y} - \widetilde{\mathbf{x}}\widehat{\boldsymbol{\beta}}_\lambda)^2 \middle| \widehat{\boldsymbol{\beta}}_\lambda\right].$$

La somme des carrés des résidus est définie par

$$SSE(\lambda) = \sum_{i=1}^{n} \left(Y_i - \mathbf{X}_{i,.} \widehat{\beta}_\lambda \right)^2 = \left\| \mathbf{Y} - \mathbf{X} \widehat{\beta}_\lambda \right\|^2,$$

où $\mathbf{x}_{i,.} \widehat{\beta}_\lambda$ est la prédiction pour l'individu i. C'est une **estimation trop optimiste de l'erreur de prédiction** sur de nouvelles données.

Pour les estimateurs ridge ou lasso, si $0 \leq \lambda_1 < \lambda_2$, on a

$$SSE(\lambda_1) < SSE(\lambda_2).$$

2.1 Mettre une observation de côté

Il s'agit de la validation croisée dite « *leave-one-out* » ou LOO. Pour chaque observation de $i = 1$ à n, on va

- ajuster le modèle sans la i-ème observation Y_i, et donc calculer un $\widehat{\boldsymbol{\beta}}_\lambda$
- utiliser le modèle, donc la valeur de $\widehat{\boldsymbol{\beta}}_\lambda$ pour prédire Y_i,
- enregistrer la valeur prédite dans $\widehat{Y}_{i,\lambda}$.

Si on veut comparer différentes valeurs de λ pour l'estimateur ridge, ou l'estimateur lasso, on peut alors comparer les erreurs

$$SSE_{\text{LOO}}(\lambda) = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}_\lambda\right\|^2.$$

Sauf quand n est petit, on emploie plutôt d'autres méthodes. En effet, cette méthode de validation croisée est

- **lente** : on doit faire autant d'estimation qu'il y a d'observations
- **peu stable** : supprimer une seule observation ne permet pas de changer beaucoup le jeu de données.

2.2 K blocs

Cette fois-ci, on commence par diviser totalement au hasard le jeu de données en K blocs de tailles comparables. Et on prépare un vecteur $\widehat{\mathbf{Y}}_\lambda$ de dimension n, pour enregister les prédictions par validation croisée sur chaque observation. Puis, pour chaque bloc de $k = 1$ à K, on va

- ajuster le modèle sans le k-ème bloc d'observations, et donc calculer $\widehat{\boldsymbol{\beta}}_\lambda$

— utiliser le modèle, donc la valeur de $\widehat{\boldsymbol{\beta}}_\lambda$ pour prédire les Y_i du k-ème bloc,
— enregistrer les valeurs prédites de Y_i dans les bonnes coordonnées de $\widehat{\mathbf{Y}}_\lambda$.

À nouveau, l'erreur de généralisation de $\widehat{\boldsymbol{\beta}}_\lambda$ se calcule avec

$$SSE_K(\lambda) = \left\|\mathbf{Y} - \widehat{\mathbf{Y}}_\lambda\right\|^2$$

En règle générale, on choisit $K = 5$ ou 10 blocs. Cette méthode de validation croisée est

— plus rapide : on ne doit faire l'estimation de $\boldsymbol{\beta}$ que K fois
— perturbe plus le jeu de données : à chaque fois, on enlève tout un bloc d'observations avant d'estimer (si $K = 5$, on enlève $\approx 20\%$ des observations)

2.3 Choisir λ

Une fois que l'on dispose d'un critère fiable, SSE_{LOO} ou SSE_K, on peut choisir λ avec

$$\widehat{\lambda} = \mathbf{argmin}_\lambda \, SSE_*(\lambda), \quad \text{où } * \in \{\text{LOO}, K\}.$$

L'estimateur de β retenu au final est donc $\widehat{\beta}_{\widehat{\lambda}}$ avec une valeur de λ calibrée sur le jeu de donnée par validation croisée.

Le problème d'optimisation de $SSE_*(\lambda)$ est résolu ainsi.

— On fixe une grille de valeurs de λ : $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_T$.

— On calcule $SSE_*(\lambda)$ en chacun des points de cette grille par validation croisée.

— On retient la valeur de λ qui donne la plus petite valeur.

Dernier problème : $SSE_*(\widehat{\lambda})$ est lui-même est une estimation trop optimiste de l'erreur de prédiction pour la valeur de λ choisie. Il faudrait donc avoir de nouvelles données pour évaluer son erreur de prédiction.

On ne peut plus utiliser la technique de validation croisée. Il faut donc garder **dès le début de l'étude** une petite partie des données comme jeu de données de **test** pour s'en servir comme « nouvelles données » à la fin.

Au final, le jeu de données est divisé est trois parties :

— **entraînement** : partie sur laquelle on ajuste les $\widehat{\beta}_\lambda$

— **validation** : partie sur laquelle on choisie λ

— **test** : partie sur laquelle on évalue l'erreur de prédication finale.

Les techniques de validation croisée permettent de regrouper entraînement et validation. Notons enfin que notre critère d'erreur ici était une différence au carré. On peut utiliser d'autres critères d'erreur (somme des valeurs absolues,. . .)